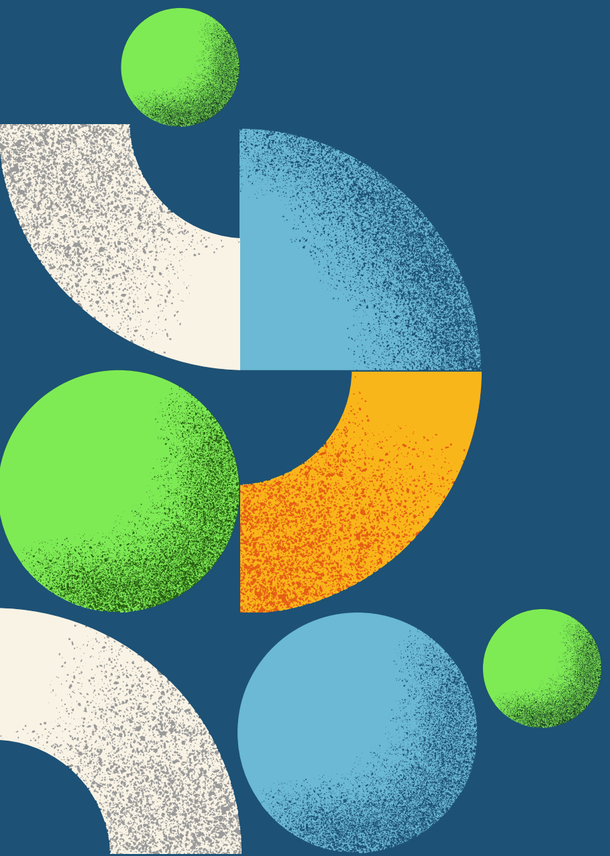
Evaluating Structured Outputs in NLP

HARDIK MITTAL (2021114016)
AYAN DATTA (2021114017)

Introduction



- **Integration & Reliability:** Structured outputs, such as JSON, serve as a bridge between unstructured natural language and deterministic downstream applications.

- **Recent advacements:** Recent advancements, including OpenAI's function calling and JSON mode, show the importance of structured output generation.

- **Reducing Post-Processing Overhead:** Adhering to predefined schemas minimizes the need for complex parsing, reducing errors.

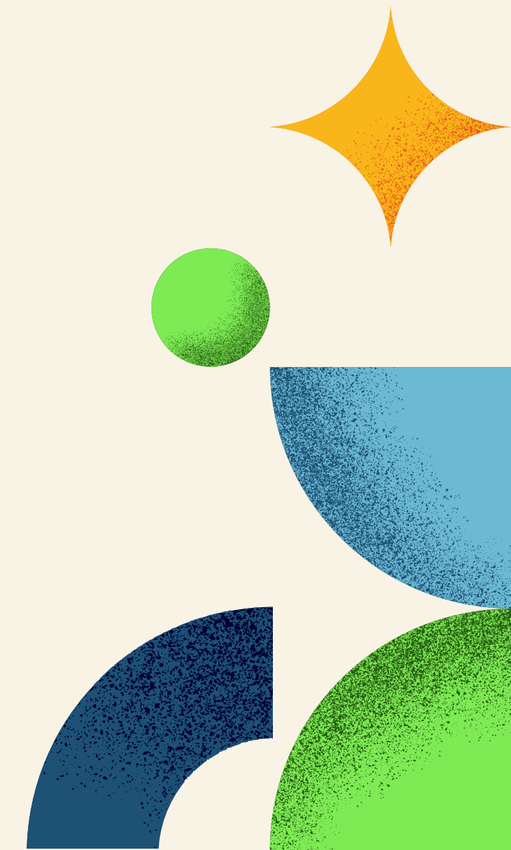
Applications and Trends

- **Real-World Usage:** APIs, chatbot systems, and enterprise data pipelines increasingly require machine-interpretable outputs to function effectively.
- **Function Calling & Schema Enforcement:** Tools require LLMs to produce outputs that conform to JSON schemas.
- **Further Directions :** This can be further expanded to other structured data like code, yaml, XML, HTML, LaTeX etc.

Problem Formalization

Evaluating generated structured outputs (e.g., JSON objects) against reference outputs to ensure accuracy and reliability.

- **Schema Correctness:** Ensuring outputs adhere to predefined schemas.
- **Content Accuracy:** Verifying that values match the ground truth, both exactly and semantically.
- **Semantic Equivalence:** Allowing flexibility in key order or formatting when the meaning remains unchanged.

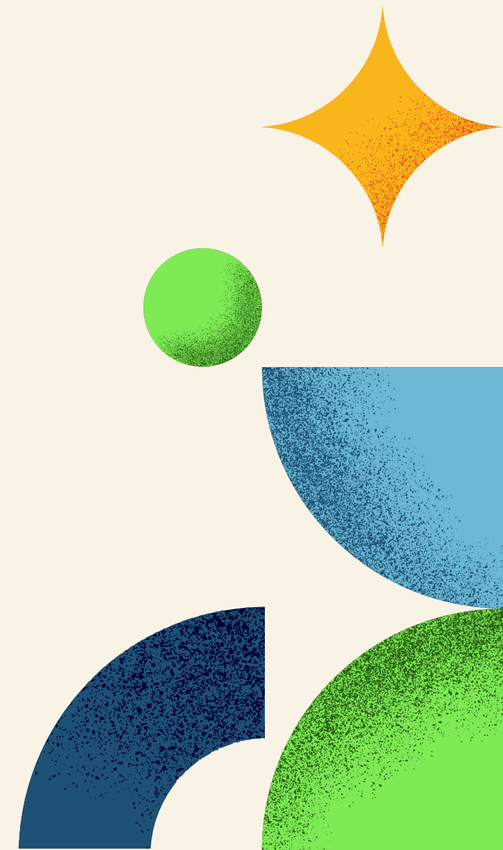


Problem Types

Evaluating generated structured outputs (e.g., JSON objects) against reference outputs to ensure accuracy and reliability.

Two types:

- **Schema-based Tasks**
  - Need output in a predefined structure
- **Schema-free Tasks**
  - No fixed structure



Schema-based

Isaac is a current undergraduate at Rutgers University NB, pursuing a double major in Economics and Information and Technology and Informatics. He hopes to pursue a career in Data Analytics or Cyber Security. He enjoys his free time diving into books and dancing. He's also learning how to cook, so any cooking advice thrown his way would be greatly appreciated.
Kelly always loved to write. She earned a Bachelor's Degree in Creative Writing/American Studies from Rutgers University. Kelly contributes as a guest writer on several blogs; her favorite topics include third wave feminism, medieval royalty, and Broadway musicals. She keeps busy on the weekends by hiking, creating mix tapes for friends and family, and teaching color guard to local high school students. Kelly resides in East Windsor, New Jersey with her husband.
Amy grew up in Somerset, New Jersey and got her degree in Journalism and Media Studies from Rutgers University. She has experience working everywhere from media marketing to the fashion industry

text

```
{
  Company: {
    Name: ,
    Description: ,
    Experience:
  },
  Team: {
    Members: [
      {
        Name: ,
        Bio: ,
        Education: ,
        Interests: []
      }
    ]
  }
}
```

template

```
{
  Company: {
    Name: ,
    Description: ,
    Experience:
  },
  Team: {
    Members: [
      {
        Name: Isaac,
        Bio: ,
        Education: Rutgers University NB,
        Interests: [diving into books, dancing, cooking]
      },
      {
        Name: Kelly,
        Bio: ,
        Education: Rutgers University,
        Interests: [
          hiking,
          creating mix tapes for friends and family,
          teaching color guard to local high school students
        ]
      },
      {
        Name: Amy,
        Bio: ,
        Education: Rutgers University,
        Interests: []
      }
    ]
  }
}
```

output

Schema-free

A mixed solution of 8 ml of water and 32 ml of dimethoxyethane was added into a flask into which 2-bromo-6-isopropylanisole (1.98 g, 8.64 mmol), phenylboronic acid (2.10 g, 17.28 mmol), palladium acetate (96 mg, 0.43 mmol), triphenylphosphine (0.225 g, 0.86 mmol), and potassium phosphate (11 g, 51.84 mmol) were already added, and then refluxed at normal temperature for 12 hours.

```
{
    reactants: [
        {name: 2-bromo-6-isopropylanisole, quantity: 1.98 g, 8.64 mmol},
        {name: phenylboronic acid, quantity: 2.10 g, 17.28 mmol},
    ],
    reagents: [
        {name: potassium phosphate, quantity: 11 g, 51.84 mmol},
        {name: water, quantity: 1.98 g, 8.64 mmol},
    ],
    solvents: [
        {name: dimethoxyethane, quantity: 32 ml}
    ],
    catalysts: [
        {name: palladium acetate, quantity: 96 mg, 0.43 mmol},
        {name: triphenylphosphine, quantity: 0.225 g, 0.86 mmol},
    ],
    time: [12 hours],
    temperature: [normal temperature],
}
```

Common Evaluation Methods & Their Critique

Schema-based

- Exact Match on values
- Traditional Metrics (BLEU, ROUGE) on corresponding values.
- BERTScore, MoverScore based metrics on corresponding values

Schema-free

- JSON Edit Distance.
- Traditional Metrics (BLEU, ROUGE) on whole JSON.
- BERTScore, MoverScore based on whole JSON.

Critique - Schema Based

Why only comparing corresponding values is a bad idea ?

Text: The dog ran towards me, barked and bit me.

Schema

```
{
    "events": [
        {"description": ""}
    ]
}
```

Reference

```
{
    "events": [
        {"description": "dog runs, and barks"},
        {"description": "dog bites"}
    ]
}
```

Prediction

```
{
    "events": [
        {"description": "dog runs"},
        {"description": "dog barks, and bites"}
    ]
}
```

Critique - Schema Based

Why only comparing corresponding values is a bad idea ? And Why the metrics are bad.

Text: The dog ran towards me, barked and bit me.

Schema

```
{
    "events": [
        {"description": ""}
    ]
}
```

Reference

```
{
    "events": [
        {"description": "dog runs, and barks"},
        {"description": "dog bites"}
    ]
}
```

Prediction

```
{
    "events": [
        {"description": "dog runs"},
        {"description": "canine woofs"},
        {"description": "doggy bites"}
    ]
}
```

Critique - Schema Free

Task: Given the wikipedia page about Apollo 8. Give details about the crew in a structured format.

Prediction


```
{
    "crew": {
        "commander": {
            "name": "Frank Borman",
            "role": "Commander",
            "background": {
                "previous_flights": [
                    "Gemini 7"
                ],
                "training": "Command Module Pilot"
            }
        },
        "command_module_pilot": {
            "name": "James Lovell",
            "role": "Command Module Pilot",
            "background": {
                "previous_flights": [
                    "Gemini 7",
                    "Gemini 12"
                ],
                "training": "Lunar Module Pilot"
            }
        }
    }
}
```

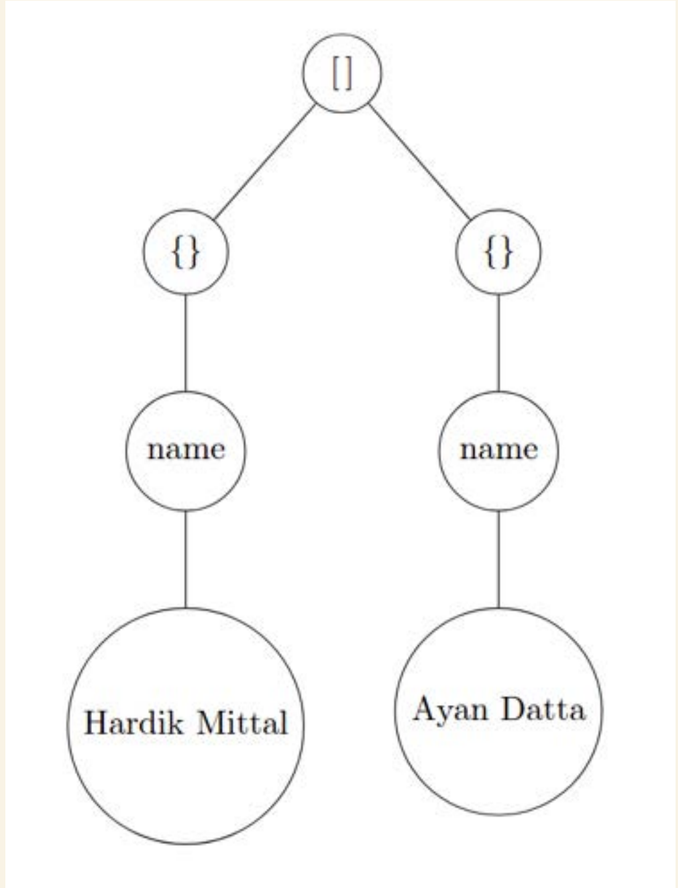

Reference


```
{
    "mission_crew": [
        {
            "position": "Commander",
            "person": {
                "full_name": "Frank Borman",
                "experience": {
                    "flights": [
                        "Gemini 7"
                    ],
                    "qualification": "Command Module Pilot"
                }
            }
        },
        {
            "position": "Command Module Pilot",
            "person": {
                "full_name": "James Lovell",
                "experience": {
                    "flights": [
                        "Gemini 7",
                        "Gemini 12"
                    ],
                    "qualification": "Lunar Module Pilot"
                }
            }
        }
    ]
}
```

Proposed New Approaches for Evaluation

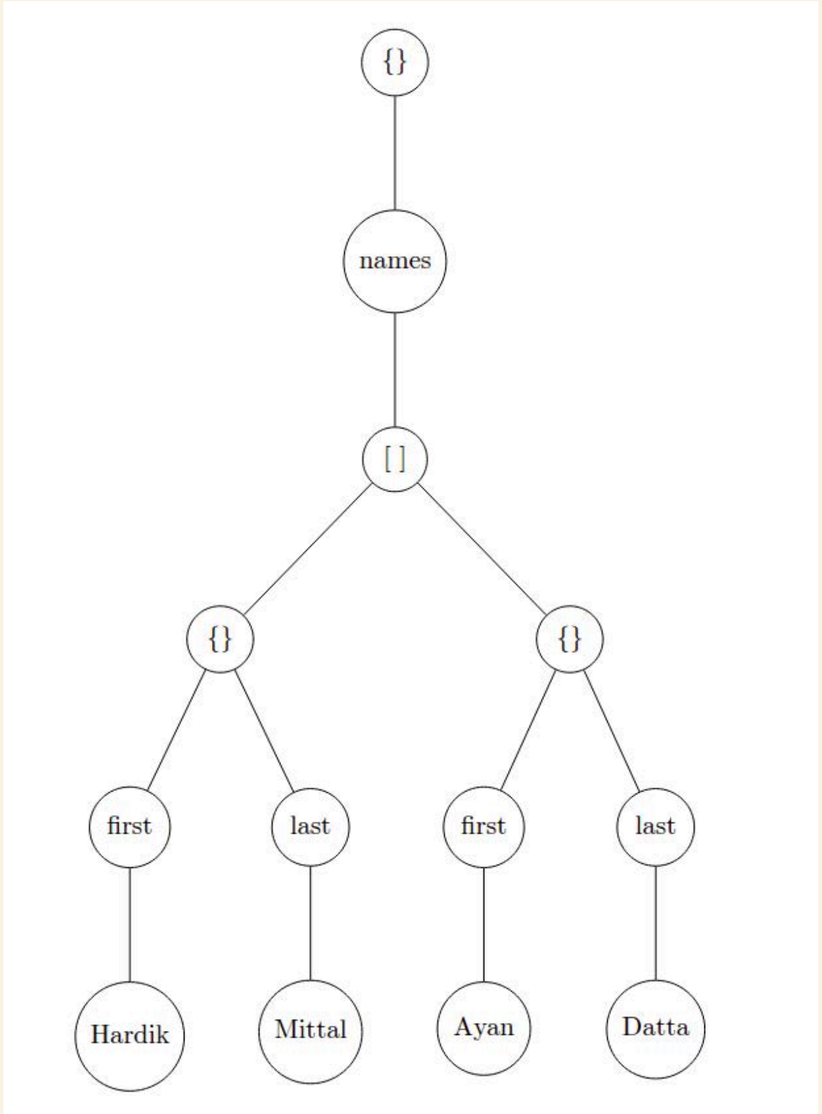
- Propose an alternate way of looking at the problem
  - Equivalent to Tree Similarity
  - Some other structured representation

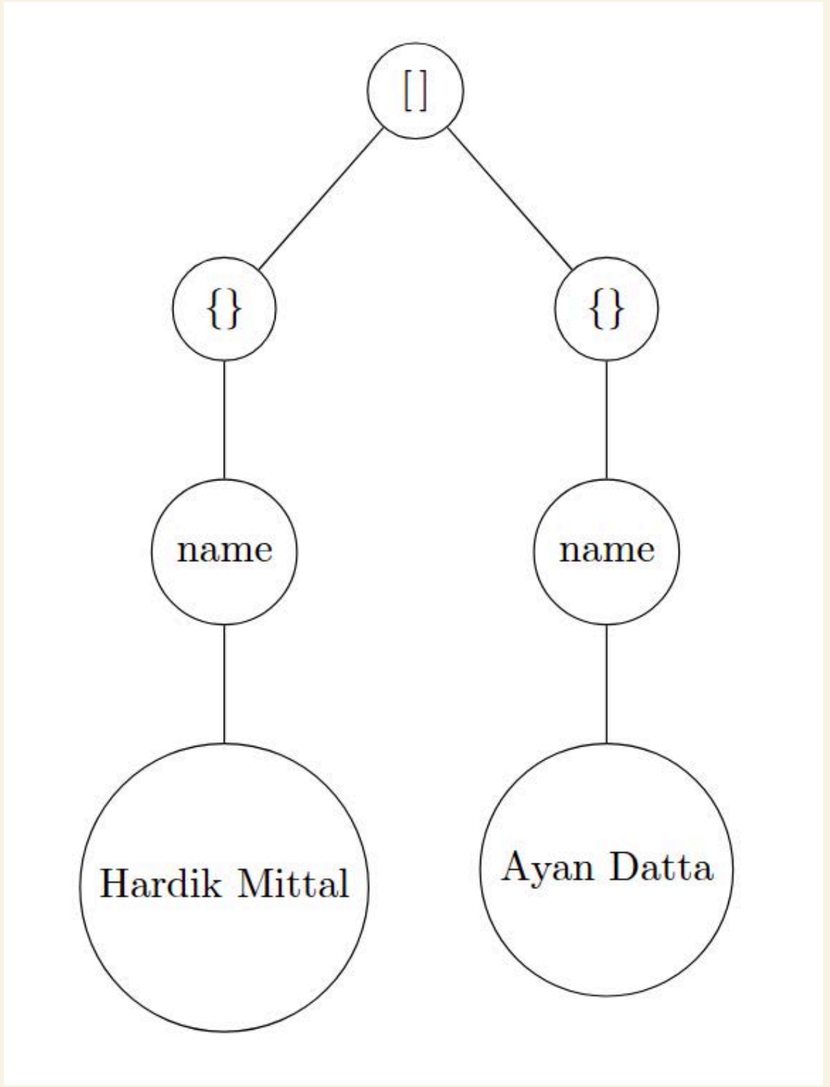
```
[
    {
        "name": "Hardik Mittal"
    },
    {
        "name": "Ayan Datta"
    }
]
```



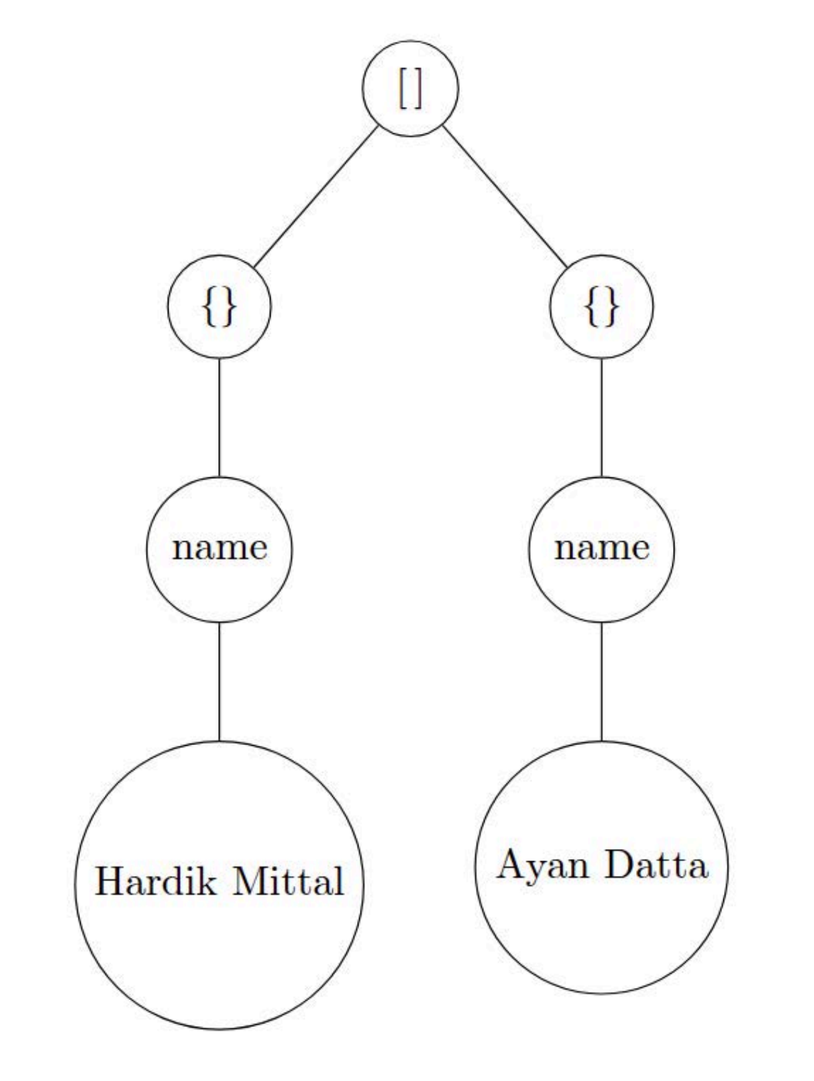
Proposed New Approaches for Evaluation

- New Equivalent Problem: How similar are these trees





Proposed New Approaches for Evaluation



- Compare Depth First Traversals using text similarity methods: Same as Previous metrics.
  (Depth First Traversal (Preorder) == JSON)
- Use Other Traversal Methods (Breadth First, Postorder, etc.)
- Node-Node Similarity - O(|V|^2)
  - ROUGE/BLEU or other statistical methods
  - Individual BERT Embeddings (BERT Scores)
  - Using rule based methods convert each path into a sentence (not necessarily grammatical), and use embeddings from that
- Edit Distance using semantic similarity.

Current Progress

- Dataset Curation:
  - No existing JSON datasets :(
  - Hard for humans to curate data
  - Scraping JSONs from the web gives very noisy data
  - Using **Qwen/Qwen2.5-14B-Instruct-GPTQ-Int8** on Wikipedia to structure information
  - Transformations to the JSON to get schema-based task outputs (take a subtree as reference for the task to extract the parent key)

- Evaluating Smaller Models like Llama-3.2-1b-Instruct, Qwen2.5-1b-Instruct

Human Evaluation - Bonus

- Hard to interpret large JSON / other structured output easily.
- **Hypothesis to test:** Is Tree form easy to comprehend **or** is the textual form easy
  - Collect Human Preferences across different structured languages

- **How easy is it for humans to evaluate which structure is better.**
  - Humans argue whether small and compact better or is big and descriptive

- What are some aspects of structured text evaluation that are easy for humans to comprehend and evaluate

Future Directions

- Reference - Free Evaluation
- Integration of Numerical Data (Maybe use stats like Mean, Variance over lists of numbers?)
- Integrating Graph Embeddings
- Study **other structured domains like:**
  - LaTeX (Document Generation Evaluation)
  - HTML
  - Code / Abstract Syntax Trees\
  - ...

Thank You